

# Stability and convergence of time discretizations of quasi-linear evolution equations of Kato type

Balázs Kovács · Christian Lubich

**Abstract** Semidiscretization in time is studied for a class of quasi-linear evolution equations in a framework due to Kato, which applies to symmetric first-order hyperbolic systems and to a variety of fluid and wave equations. In the regime where the solution is sufficiently regular, we show stability and optimal-order convergence of the linearly implicit and fully implicit midpoint rules and of higher-order implicit Runge–Kutta methods that are algebraically stable and coercive, such as the collocation methods at Gauss nodes.

**Keywords** Quasi-linear evolution equations, symmetric hyperbolic systems, dispersive equations, implicit midpoint rule, implicit Runge–Kutta method, algebraic stability, coercivity, energy estimates, error bounds.

**Mathematics Subject Classification (2010)** 65M12

## 1 Introduction

In a very insightful paper published in 1975, Kato [13] presents a concise framework for quasi-linear evolution equations in a Banach space, proves local well-posedness of the initial value problem within this framework and shows that the framework and results apply to a variety of quasi-linear partial differential equations. He lists symmetric hyperbolic systems of the first order, wave equations, Korteweg–de Vries equation, Navier–Stokes and Euler equations, equations for compressible fluids, magnetohydrodynamic equations, coupled Maxwell and Dirac equations — and adds “etc.”. Particularly noteworthy appears the application to symmetric hyperbolic systems in the sense of Friedrichs (in arbitrary space dimension), which is a large and fundamental class of problems.

---

Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle, D-72076 Tübingen, Germany.  
E-mail: {Kovacs,Lubich}@na.uni-tuebingen.de

While Kato’s paper has been influential and highly cited in the analysis of nonlinear hyperbolic and dispersive partial differential equations, it has apparently gone unnoticed in the numerical literature for such equations. Kato’s framework has been modified and generalized to further classes of partial differential equations, by himself and coauthors in [11, 12] shortly after [13], and by other researchers until recently, e.g., in [7, 17]. To our knowledge, the only numerical paper related to Kato’s framework is the recent work of Hochbruck & Pažur [10] who study the implicit Euler time discretization in a modified Kato framework that was developed by Müller [17] for dealing with a class of quasi-linear Maxwell equations. We acknowledge that it was [10] and [17] that led us to the present work.

Here we show that Kato’s original framework from [13], when restricted to Hilbert spaces (which are mostly used in the applications), combines remarkably well with the technique of “energy estimates” for time discretizations, that is, with the use of positive definite and semi-definite bilinear forms for proving stability and error bounds. We show this first for the implicit midpoint rule in Section 3, and then (in Section 4) for implicit Runge–Kutta methods such as the Gauss and Radau IIA methods of arbitrary orders, which have the properties of algebraic stability and coercivity, notions that are due to Burrage & Butcher [1] and Crouzeix [4] (for algebraic stability) and to Crouzeix and Raviart [5] (for coercivity); see also [6, 9]. Although these notions were developed and recognized as important properties in the context of stiff ordinary differential equations in the same decade in which Kato’s paper appeared, it seems that no link between these analytical and numerical theories was made. With a delay of some decades, this is now done in the present paper — in view of both, the perfectly fitting connection of the analytical framework and the numerical methods, and the undiminished significance of the considered evolution equations in applications.

We study only time discretization in this paper. Effects of truncation of an unbounded spatial domain and of space discretization are not considered here. Moreover, we work in a regime where a sufficiently regular solution exists. Of course, we are aware that shocks may develop in finite time in quasi-linear hyperbolic equations. Nevertheless, for many cases within the class of evolution equations considered (in particular, in problems of wave propagation and dispersive equations), regular solutions exist for sufficiently long times of interest (or even for all times), and it is then important to understand the mechanisms that yield stability and convergence of numerical discretizations.

## 2 Kato’s framework in a Hilbert space setting

We consider a quasi-linear evolution equation (with  $\dot{\phantom{x}} = d/dt$ )

$$\dot{u} + A(u)u = f(u) \tag{2.1}$$

in the following setting, which is a Hilbert space version of the framework in Kato’s paper [13]. Let  $X$  and  $Y$  be two real Hilbert spaces such that  $Y$  is

densely and continuously embedded in  $X$ . We denote the inner product on  $X$  by  $(\cdot, \cdot)$  and the norms on  $X$  and  $Y$  by

$$|\cdot| = \|\cdot\|_X, \quad \|\cdot\| = \|\cdot\|_Y.$$

For convenience we choose the norms such that  $|y| \leq \|y\|$  for all  $y \in Y$ . We assume throughout this paper that for every  $R > 0$  the following assumptions are satisfied, with real numbers  $M_R^A, M_R^B, L_R^A, L_R^B, \ell_R^X, \ell_R^Y$  depending only on  $R$ :

(K1) (*m-accretivity* [14, Section V.10]) For every  $y \in Y$ , the closed linear operator  $A(y)$  on  $X$  has the open left complex half-plane in the resolvent set and satisfies the bound

$$(w, A(y)w) \geq 0 \quad \text{for all } w \in D(A(y)). \quad (2.2)$$

Moreover, the domain  $D(A(y))$  contains the space  $Y$ , and there is the  $Y$ -locally uniform bound, for  $y \in Y$  with  $\|y\| \leq R$ ,

$$|A(y)w| \leq M_R^A \|w\| \quad \text{for all } w \in Y. \quad (2.3)$$

(K2) (*Kato's commutator condition*) There exists an isometry  $S : Y \rightarrow X$ , self-adjoint as a linear operator on  $X$ , with the following property: For every  $y \in Y$  with  $\|y\| \leq R$ ,

$$SA(y)S^{-1} = A(y) + B(y) \quad (2.4)$$

(with equality of domains), where  $B(y)$  is  $Y$ -locally uniformly bounded on  $X$ :

$$|B(y)v| \leq M_R^B |v| \quad \text{for all } v \in X. \quad (2.5)$$

(K3) (*Y-local Lipschitz conditions*) For all  $y, \tilde{y} \in Y$  with  $\|y\| \leq R, \|\tilde{y}\| \leq R$ ,

$$|(A(y) - A(\tilde{y}))w| \leq L_R^A \|y - \tilde{y}\| \|w\| \quad \text{for all } w \in Y, \quad (2.6)$$

$$|(B(y) - B(\tilde{y}))v| \leq L_R^B \|y - \tilde{y}\| |v| \quad \text{for all } v \in X. \quad (2.7)$$

(K4) (*Semilinear term*) The function  $f : X \rightarrow X$  is  $Y$ -locally Lipschitz-continuous in  $X$  and  $Y$ : For all  $y, \tilde{y} \in Y$  with  $\|y\| \leq R, \|\tilde{y}\| \leq R$ ,

$$|f(y) - f(\tilde{y})| \leq \ell_R^X |y - \tilde{y}|, \quad (2.8)$$

$$\|f(y) - f(\tilde{y})\| \leq \ell_R^Y \|y - \tilde{y}\|. \quad (2.9)$$

In [13], Kato just assumes Banach spaces instead of Hilbert spaces, and he requires that  $-A(y)$  is the generator of a contraction semigroup on  $X$ . On a Hilbert space, this condition is equivalent to (K1) by the Lumer–Phillips theorem [18, p. 14].

Under these conditions, Kato [13, Theorem 6] proves local existence and uniqueness of a solution to (2.1) in  $C([0, \bar{t}]; Y) \cap C^1([0, \bar{t}]; X)$  (for some  $\bar{t} > 0$ ) for initial data in  $Y$ .

He then proceeds to show that (K1)–(K4) are indeed satisfied for a wide variety of quasi-linear partial differential equations, as listed in the introduction. In these applications, he has typically

$$X = L_2(\mathbb{R}^d), \quad Y = H^s(\mathbb{R}^d), \quad \text{and the isometry } S = (I - \Delta)^{s/2} \quad (2.10)$$

for an exponent  $s > 0$  that is sufficiently large so that the  $s$ th-order Sobolev space  $H^s$  is a Banach algebra ( $s > d/2$ ).

Moreover, Kato [13, Theorem 7] also gives a perturbation result. Here we give another perturbation result together with its simple proof, because its time-discrete versions will be important later in this paper.

Suppose  $u(t) \in Y$  solves (2.1) for  $0 \leq t \leq T$  and  $u^*(t) \in Y$  solves (2.1) up to a defect  $d(t) \in Y$  for  $0 \leq t \leq T$ :

$$\dot{u}^* + A(u^*)u^* = f(u^*) + d. \quad (2.11)$$

**Lemma 2.1** *In the above situation, suppose that, for  $0 \leq t \leq T$ ,*

$$\|u^*(t)\| \leq R \quad \text{and} \quad Su^*(t) \in Y \quad \text{with} \quad \|Su^*(t)\| \leq K.$$

*Then, there exists  $\delta > 0$ , which depends only on  $K$ ,  $R$  and  $T$  such that for perturbations satisfying*

$$\|u(0) - u^*(0)\|^2 + \int_0^T \|d(s)\|^2 ds \leq \delta^2,$$

*the error  $u - u^*$  satisfies, for  $0 \leq t \leq T$ ,*

$$\begin{aligned} \|u(t) - u^*(t)\|^2 &\leq C_Y \left( \|u(0) - u^*(0)\|^2 + \int_0^t \|d(s)\|^2 ds \right), \\ |u(t) - u^*(t)|^2 &\leq C_X \left( |u(0) - u^*(0)|^2 + \int_0^t |d(s)|^2 ds \right), \end{aligned}$$

*where  $C_Y$  depends only on  $K$ ,  $R$  and  $T$ , and  $C_X$  depends only on  $R$  and  $T$ .*

*Remark* In the situation (2.10) the condition  $Su^*(t) \in Y$  means  $u^*(t) \in H^{2s}$ . This higher regularity is again ensured locally in time for initial values in  $H^{2s}$ , using Kato's theory with  $2s$  in place of  $s$ .

*Proof* The error  $e = u - u^*$  satisfies the error equation

$$\dot{e} + A(u)e = -(A(u) - A(u^*))u^* + (f(u) - f(u^*)) - d. \quad (2.12)$$

Using (2.4) as  $A(y) = S^{-1}A(y)S + S^{-1}B(y)S$ , this becomes

$$\begin{aligned} \dot{e} + S^{-1}A(u)Se + S^{-1}B(u)Se \\ = -S^{-1}(A(u) - A(u^*))Su^* - S^{-1}(B(u) - B(u^*))Su^* + (f(u) - f(u^*)) - d. \end{aligned}$$

On applying the operator  $S$  on both sides we thus have

$$\begin{aligned} & S\dot{e} + A(u)Se + B(u)Se \\ &= -(A(u) - A(u^*))Su^* - (B(u) - B(u^*))Su^* + (Sf(u) - Sf(u^*)) - Sd. \end{aligned}$$

Using the accretivity (2.2) and the bounds (2.5)–(2.9) and recalling that  $S$  is an isometry between  $Y$  and  $X$ , we therefore obtain, as long as  $\|u(t)\| \leq 2R$ ,

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|e\|^2 &= \frac{1}{2} \frac{d}{dt} |Se|^2 = (Se, S\dot{e}) \\ &\leq M_{2R}^B \|e\|^2 + L_{2R}^A \|e\| |e| \|Su^*\| + L_{2R}^B \|e\|^2 |Su^*| + \ell_{2R}^Y \|e\|^2 + \|e\| \|d\| \\ &\leq (M_{2R}^B + L_{2R}^A K + L_{2R}^B R + \ell_{2R}^Y + \frac{1}{2T}) \|e\|^2 + \frac{T}{2} \|d\|^2, \end{aligned}$$

and the error bound in the  $Y$ -norm follows with Gronwall's inequality. Choosing  $\delta$  so small that  $C_Y \delta \leq R$ , the condition  $\|u(t)\| \leq 2R$  then remains satisfied for  $0 \leq t \leq T$ . Taking in (2.12) the inner product with  $e$  and using (2.2), (2.6), (2.8) gives us

$$\frac{1}{2} \frac{d}{dt} |e|^2 = (e, \dot{e}) \leq (L_{2R}^A R + \ell_{2R}^X) |e|^2 + |e| |d|,$$

and finally the Gronwall inequality yields the error bound in the  $X$ -norm.  $\square$

### 3 Linearly implicit and fully implicit midpoint rules

For the time discretization of (2.1) we first consider variants of the implicit midpoint rule. For a positive stepsize  $\tau$  and integers  $n = 0, 1, 2, \dots$ , the solution  $u(t)$  to (2.1) with initial value  $u_0$  is approximated at  $t_n = n\tau$  by  $u_n$ , which is determined by

$$\frac{u_{n+1} - u_n}{\tau} + A(\hat{u}_{n+1/2}) \frac{u_{n+1} + u_n}{2} = f(\hat{u}_{n+1/2}). \quad (3.1)$$

Here we set either

- (FI)  $\hat{u}_{n+1/2} = \frac{u_{n+1} + u_n}{2}$  for the fully implicit midpoint rule, or
- (LI)  $\hat{u}_{n+1/2} = u_n + \frac{1}{2}(u_n - u_{n-1})$  for a linearly implicit midpoint rule.

In the latter case, we set  $\hat{u}_{1/2} = u_0$  in the first step.

For ease of presentation, we just consider constant stepsizes in this paper, but all our results generalize to variable stepsizes (with a bounded ratio of subsequent stepsizes) without any additional difficulty.

#### 3.1 Stability of the linearly implicit midpoint rule

We begin with the stability analysis of the linearly implicit method, (3.1) with (LI) in the setting of Section 2. As is clear from the framework of Section 2, it is important to control the  $Y$ -norm of the numerical solution. This can be done *per se* (Lemma 3.1) or in comparison with the exact solution (Lemma 3.2).

**Lemma 3.1** *Suppose that for all  $k \leq n$  we have  $u_k \in Y$  with  $\|u_k\| \leq R$ . Then there exist  $\tau_R > 0$  and  $C_R, \gamma_R \geq 0$ , which depend only on  $R$  through the constants in conditions (K1)–(K4), such that for stepsizes  $\tau \leq \tau_R$  the linearly implicit midpoint rule (3.1) with (LI) has a unique solution  $u_{n+1} \in Y$ . Moreover, this is bounded by*

$$\|u_{n+1}\| \leq (1 + C_R \tau) \|u_n\| + \tau \gamma_R.$$

*Proof* Let us introduce the abbreviations

$$u_{n+1/2} = \frac{u_{n+1} + u_n}{2}, \quad \dot{u}_{n+1/2} = \frac{u_{n+1} - u_n}{\tau}, \quad (3.2)$$

so that the numerical method (3.1) reads more concisely

$$\dot{u}_{n+1/2} + A(\hat{u}_{n+1/2})u_{n+1/2} = f(\hat{u}_{n+1/2}).$$

With (2.4), this is written equivalently as

$$\dot{u}_{n+1/2} + S^{-1}A(\hat{u}_{n+1/2})Su_{n+1/2} + S^{-1}B(\hat{u}_{n+1/2})Su_{n+1/2} = f(\hat{u}_{n+1/2}),$$

where we note that  $\|\hat{u}_{n+1/2}\| = \|\frac{3}{2}u_n - \frac{1}{2}u_{n-1}\| \leq 2R$ . We apply  $S$  to both sides of the equation and obtain a linear equation in  $X$  for  $Su_{n+1}$  with the operator  $\tau^{-1}I + A(\hat{u}_{n+1/2}) + B(\hat{u}_{n+1/2})$ , which by (K1) is invertible if  $1/\tau > \|B(\hat{u}_{n+1/2})\|$ . In view of (2.5) this is satisfied if  $\tau M_{2R}^B < 1$ . Hence, under this stepsize restriction we have a unique solution  $u_{n+1} \in Y$ .

To derive the bound for  $\|u_{n+1}\|$ , we take the inner product with  $Su_{n+1/2}$  in the equation. With the accretivity (2.2) and the bound (2.5) we obtain

$$(Su_{n+1/2}, S\dot{u}_{n+1/2}) \leq M_{2R}^B |Su_{n+1/2}|^2 + |Su_{n+1/2}| |Sf(\hat{u}_{n+1/2})|.$$

The term on the left-hand side is

$$(Su_{n+1/2}, S\dot{u}_{n+1/2}) = \frac{1}{2\tau} (|Su_{n+1}|^2 - |Su_n|^2) = \frac{1}{2\tau} (\|u_{n+1}\|^2 - \|u_n\|^2).$$

On the right-hand side we note

$$\begin{aligned} |Sf(\hat{u}_{n+1/2})| &= \|f(\hat{u}_{n+1/2})\| \leq \|f(\hat{u}_{n+1/2}) - f(0)\| + \|f(0)\| \\ &\leq \ell_{2R}^Y 2R + \|f(0)\| =: c_{2R}. \end{aligned}$$

Hence we obtain

$$\|u_{n+1}\|^2 \leq \|u_n\|^2 + \tau M_{2R}^B \|u_{n+1/2}\|^2 + \tau c_{2R} \|u_{n+1/2}\|.$$

Using that  $\|u_{n+1/2}\|^2 \leq \frac{1}{2}(\|u_{n+1}\|^2 + \|u_n\|^2)$ , the result follows.  $\square$

Suppose that  $u_n \in Y$  solves (3.1) with (LI) for  $0 \leq n\tau \leq T$ , and  $u_n^* \in Y$  (which will later be taken as the exact solution value  $u(t_n)$ ) solves (3.1) with (LI) up to a defect  $d_{n+1/2} \in Y$  for  $0 \leq n\tau \leq T$ :

$$\frac{u_{n+1}^* - u_n^*}{\tau} + A(\hat{u}_{n+1/2}^*) \frac{u_{n+1}^* + u_n^*}{2} = f(\hat{u}_{n+1/2}^*) + d_{n+1/2}, \quad (3.3)$$

with  $\hat{u}_{n+1/2}^* = u_n^* + \frac{1}{2}(u_n^* - u_{n-1}^*)$ .

We then have the following time-discrete version of Lemma 2.1.

**Lemma 3.2** *In the above situation, suppose that, for  $0 \leq n\tau \leq T$ ,*

$$\|u_n^*\| \leq R \quad \text{and} \quad Su_n^* \in Y \quad \text{with} \quad \|Su_n^*\| \leq K.$$

*Then, there exist  $\bar{\tau} > 0$  and  $\delta > 0$ , which depend only on  $K$ ,  $R$  and  $T$ , such that for stepsizes  $\tau \leq \bar{\tau}$  and perturbations satisfying*

$$\|u_0 - u_0^*\|^2 + \tau \sum_{0 \leq n\tau \leq T} \|d_{n+1/2}\|^2 \leq \delta^2,$$

*the error satisfies, for  $0 \leq n\tau \leq T$ ,*

$$\begin{aligned} \|u_n - u_n^*\|^2 &\leq C_Y \left( \|u_0 - u_0^*\|^2 + \tau \sum_{k=0}^{n-1} \|d_{k+1/2}\|^2 \right), \\ |u_n - u_n^*|^2 &\leq C_X \left( |u_0 - u_0^*|^2 + \tau \sum_{k=0}^{n-1} |d_{k+1/2}|^2 \right), \end{aligned}$$

where  $C_Y$  depends only on  $K$ ,  $R$  and  $T$ , and  $C_X$  depends only on  $R$  and  $T$ .

*Proof* The proof transfers the arguments of the proof of Lemma 2.1 to the discrete case. With the error  $e_n = u_n - u_n^*$  we associate the abbreviations (cf. (3.2))

$$e_{n+1/2} = \frac{e_{n+1} + e_n}{2}, \quad \dot{e}_{n+1/2} = \frac{e_{n+1} - e_n}{\tau}, \quad \widehat{e}_{n+1/2} = e_n + \frac{1}{2}(e_n - e_{n-1}).$$

We have the error equation

$$\begin{aligned} \dot{e}_{n+1/2} + A(\widehat{u}_{n+1/2})e_{n+1/2} &= - (A(\widehat{u}_{n+1/2}) - A(\widehat{u}_{n+1/2}^*))u_{n+1/2}^* \\ &\quad + (f(\widehat{u}_{n+1/2}) - f(\widehat{u}_{n+1/2}^*)) - d_{n+1/2}. \end{aligned} \quad (3.4)$$

Using (2.4) as  $A(y) = S^{-1}A(y)S + S^{-1}B(y)S$  and applying the operator  $S$  on both sides we thus have

$$\begin{aligned} S\dot{e}_{n+1/2} + A(\widehat{u}_{n+1/2})Se_{n+1/2} + B(\widehat{u}_{n+1/2})Se_{n+1/2} \\ = (A(\widehat{u}_{n+1/2}) - A(\widehat{u}_{n+1/2}^*))Su_{n+1/2}^* + (B(\widehat{u}_{n+1/2}) - B(\widehat{u}_{n+1/2}^*))Su_{n+1/2}^* \\ + (Sf(\widehat{u}_{n+1/2}) - Sf(\widehat{u}_{n+1/2}^*)) - Sd_{n+1/2}. \end{aligned}$$

Using the accretivity (2.2) and the bounds (2.5)–(2.9) and recalling that  $S$  is an isometry between  $Y$  and  $X$ , we therefore obtain, as long as  $\|u_n\| \leq 2R$ ,

$$\begin{aligned} \frac{1}{2\tau} (\|e_{n+1}\|^2 - \|e_n\|^2) &= \frac{1}{2\tau} (|Se_{n+1}|^2 - |Se_n|^2) = (Se_{n+1/2}, S\dot{e}_{n-1/2}) \\ &\leq M_{2R}^B \|e_{n+1/2}\|^2 \\ &\quad + L_{2R}^A \|e_{n+1/2}\| |\widehat{e}_{n+1/2}| \|Su_{n+1/2}^*\| + L_{2R}^B \|e_{n+1/2}\| |\widehat{e}_{n+1/2}| |Su_{n+1/2}^*| \\ &\quad + \ell_{2R}^Y \|e_{n+1/2}\| |\widehat{e}_{n+1/2}| + \|e_{n+1/2}\| \|d_{n+1/2}\|. \end{aligned}$$

The right-hand side is bounded by  $C_{K,R}(\|e_{n+1}\|^2 + \|e_n\|^2 + \|e_{n-1}\|^2) + \|d_{n+1/2}\|^2$  and the stated error bound in the  $Y$ -norm then follows on summing up and

using a discrete Gronwall inequality. Choosing  $\delta$  such that  $C_Y \delta \leq R$ , the condition  $\|u_n\| \leq 2R$  then remains satisfied for  $0 \leq n\tau \leq T$ . Taking in (3.4) the inner product with  $e_{n+1/2}$  and using (2.2), (2.6), (2.8) gives us

$$\frac{1}{2\tau}(|e_{n+1}|^2 - |e_n|^2) = (e_{n+1/2}, \dot{e}_{n+1/2}) \leq C_R(|e_{n+1}|^2 + |e_n|^2 + |e_{n-1}|^2) + |d_{n+1/2}|^2,$$

and finally a discrete Gronwall inequality yields the stated error bound in the  $X$ -norm.  $\square$

### 3.2 Existence and stability for the fully implicit midpoint rule

**Lemma 3.3** *The statement of Lemma 3.1 is also valid for the fully implicit midpoint rule, (3.1) with (FI).*

*Proof* The proof transfers the existence proof for (2.1) in [13] to the time discretization. We consider the fixed-point iteration, with starting iterate  $u_{n+1/2}^{(0)} = u_n$ ,

$$\frac{u_{n+1/2}^{(k+1)} - u_n}{\tau/2} + A(u_{n+1/2}^{(k)})u_{n+1/2}^{(k+1)} = f(u_{n+1/2}^{(k)}).$$

If this iteration converges to a limit  $u_{n+1/2}$ , then  $u_{n+1} = 2u_{n+1/2} - u_n$  solves (3.1) with (FI). We write the above iteration briefly as

$$u_{n+1/2}^{(k+1)} = \Phi(u_{n+1/2}^{(k)}).$$

Let  $B_{2R} := \{y \in Y : \|y\| \leq 2R\}$ , which is a closed set in  $X$ , as is stated (without proof) in [13, Lemma 7.3]. [This follows from a duality and density argument: for  $y \in Y$ ,  $\|y\| = \sup_{0 \neq v \in X} (y, v) / \|v\|_*$ , where  $\|\cdot\|_*$  is the norm on the dual  $Y'$  and we use the Gelfand triple  $Y \subset X \subset Y'$  with dense and continuous embeddings. With this formula for  $\|y\|$  it follows that for every sequence  $(y_n)$  in  $B_{2R}$  that converges to  $x \in X$  in the  $X$ -norm, also  $x \in B_{2R}$ .] Therefore,  $B_{2R}$  is a complete metric space with the metric  $d(v, w) = |v - w|$ .

By the argument of the proof of Lemma 3.1 we find that for all  $v \in B_{2R}$ ,

$$\|\Phi(v)\| \leq (1 + C_R \tau) \|u_n\| + \tau \gamma_R \leq 2R$$

for sufficiently small stepsize  $\tau \leq \tau_R$ . Hence,  $\Phi$  maps  $B_{2R}$  into itself.

We now show that  $\Phi$  is a contraction on  $B_{2R}$  for sufficiently small  $\tau$ . For  $v, \tilde{v} \in B_{2R}$ , let  $w = \Phi(v)$  and  $\tilde{w} = \Phi(\tilde{v})$ . Then,

$$\frac{\tilde{w} - w}{\tau/2} + A(\tilde{v})(\tilde{w} - w) = -(A(\tilde{v}) - A(v))w + f(\tilde{v}) - f(v).$$

Taking the inner product with  $\tilde{w} - w$  and using conditions (K1)–(K4), we obtain

$$\frac{2}{\tau} |\tilde{w} - w|^2 \leq |\tilde{w} - w| (L_{2R}^A \cdot 2R + \ell_{2R}^X) |\tilde{v} - v|$$



and hence

$$|\Phi(\tilde{v}) - \Phi(v)| = |\tilde{w} - w| \leq c_R \tau |\tilde{v} - v|.$$

Therefore, if  $c_R \tau < 1$ , then  $\Phi$  is a contraction on  $B_{2R}$ , and the result follows with the Banach fixed-point theorem.  $\square$

**Lemma 3.4** *The statement of Lemma 3.2 is also valid for the fully implicit midpoint rule, (3.1) with (FI).*

*Proof* With Lemma 3.3 at hand, the result follows with the proof of Lemma 3.2.  $\square$

### 3.3 Consistency error

We now choose the exact solution values  $u_n^* = u(t_n)$  in (3.3), with  $\hat{u}_{n+1/2}^* = u_n^* + \frac{1}{2}(u_n^* - u_{n-1}^*)$  in the case of the linearly implicit midpoint rule (except for  $n = 0$ , where  $\hat{u}_{1/2}^* = u(0)$ ), and with  $\hat{u}_{n+1/2}^* = u_{n+1/2}^* = \frac{1}{2}(u_{n+1}^* + u_n^*)$  in the case of the fully implicit midpoint rule. The defects  $d_{n+1/2}$  in (3.3) are then the consistency errors and are bounded as follows.

**Lemma 3.5** *Suppose that the exact solution  $u$  of (2.1) has the regularity  $u \in C^3([0, T], Y)$  with  $Su \in C^2([0, T], Y)$ . Then, the consistency errors (3.3) of the linearly and fully implicit midpoint rule are bounded by*

$$\|d_{n+1/2}\| \leq C\tau^2,$$

where  $C$  is independent of  $n$  and  $\tau$  with  $0 \leq n\tau \leq T - \tau$  (except for  $n = 0$  for the linearly implicit method, where  $\|d_{1/2}\| \leq C\tau$ ).

*Proof* First we note that Taylor expansion of  $u$  at  $t_{n+1/2} = (n + 1/2)\tau$  yields

$$\begin{aligned} \|\hat{u}_{n+1/2}^* - u(t_{n+1/2})\| &\leq c \max_{t_n \leq t \leq t_{n+1}} \|\ddot{u}(t)\| \cdot \tau^2, \\ \|Su_{n+1/2}^* - Su(t_{n+1/2})\| &\leq c \max_{t_{n-1} \leq t \leq t_{n+1}} \|S\ddot{u}(t)\| \cdot \tau^2, \\ \left\| \frac{u(t_{n+1}) - u(t_n)}{\tau} - \dot{u}(t_{n+1/2}) \right\| &\leq c \max_{t_n \leq t \leq t_{n+1}} \|\ddot{u}(t)\| \cdot \tau^2. \end{aligned}$$

We denote

$$\begin{aligned} r_{n+1/2} &:= A(\hat{u}_{n+1/2}^*)u_{n+1/2}^* - A(u(t_{n+1/2}))u(t_{n+1/2}) \\ &= A(\hat{u}_{n+1/2}^*)(u_{n+1/2}^* - u(t_{n+1/2})) + (A(\hat{u}_{n+1/2}^*) - A(u(t_{n+1/2})))u(t_{n+1/2}). \end{aligned}$$

Using (2.4), this becomes

$$\begin{aligned} r_{n+1/2} &= S^{-1}A(\hat{u}_{n+1/2}^*)S(u_{n+1/2}^* - u(t_{n+1/2})) \\ &\quad + S^{-1}B(\hat{u}_{n+1/2}^*)S(u_{n+1/2}^* - u(t_{n+1/2})) \\ &\quad + S^{-1}(A(\hat{u}_{n+1/2}^*) - A(u(t_{n+1/2})))Su(t_{n+1/2}) \\ &\quad + S^{-1}(B(\hat{u}_{n+1/2}^*) - B(u(t_{n+1/2})))Su(t_{n+1/2}). \end{aligned}$$

By conditions (K1)–(K3) with  $R = \max_{0 \leq t \leq T} \|u(t)\|$ , this is bounded by

$$\begin{aligned} \|r_{n+1/2}\| &\leq M_R^A \|Su_{n+1/2}^* - Su(t_{n+1/2})\| + M_R^B |Su_{n+1/2}^* - Su(t_{n+1/2})| \\ &\quad + L_R^A |\hat{u}_{n+1/2}^* - u(t_{n+1/2})| \|Su(t_{n+1/2})\| \\ &\quad + L_R^B \|\hat{u}_{n+1/2}^* - u(t_{n+1/2})\| |Su(t_{n+1/2})|. \end{aligned}$$

Moreover,  $\|f(\hat{u}_{n+1/2}^*) - f(u(t_{n+1/2}))\| \leq \ell_R^Y \|\hat{u}_{n+1/2}^* - u(t_{n+1/2})\|$ . Since

$$\begin{aligned} d_{n+1/2} &= \left( \frac{u(t_{n+1}) - u(t_n)}{\tau} - \dot{u}(t_{n+1/2}) \right) - r_{n+1/2} \\ &\quad - (f(\hat{u}_{n+1/2}^*) - f(u(t_{n+1/2}))), \end{aligned}$$

the result follows with the above estimates.  $\square$

### 3.4 Error bounds

Combining the lemmas of this section, we obtain the following error bound.

**Theorem 3.1** *Let the conditions (K1)–(K4) be satisfied, and suppose that the solution  $u$  of (2.1) has the regularity  $u \in C^3([0, T], Y)$  with  $Su \in C^2([0, T], Y)$ . Then, there exists  $\bar{\tau} > 0$  such that for stepsizes  $0 < \tau \leq \bar{\tau}$ , the errors of the fully and linearly implicit midpoint rules (3.1) with (FI) and (LI), respectively, are bounded by*

$$\|u_n - u(t_n)\| \leq C\tau^2,$$

where  $C$  is independent of  $n$  and  $\tau$  with  $0 \leq n\tau \leq T$ .

## 4 Implicit Runge–Kutta methods

### 4.1 Method formulation and properties

For a given stepsize  $\tau > 0$ , an  $m$ -stage implicit Runge–Kutta method applied to the quasi-linear equation (2.1) determines solution approximations  $u_n \approx u(t_n)$  and internal stages  $U_{ni}$  by the equations<sup>1</sup>

$$U_{ni} = u_n + \tau \sum_{j=1}^m a_{ij} \dot{U}_{nj}, \quad i = 1, \dots, m, \quad (4.1)$$

$$u_{n+1} = u_n + \tau \sum_{i=1}^m b_i \dot{U}_{ni}, \quad (4.2)$$

where

$$\dot{U}_{ni} + A(U_{ni})U_{ni} = f(U_{ni}), \quad i = 1, \dots, m. \quad (4.3)$$

---

<sup>1</sup> Here the dot is just a suggestive notation, not a time derivative.

In the following we consider the equation without a semilinear term ( $f = 0$ ) for ease of presentation, since the semilinear term causes no substantial problems in the analysis but just leads to longer formulas. As in the previous section, all results are however readily generalized to a semilinear term satisfying (K4).

The method is determined by its coefficient matrix  $\mathcal{Q} = (a_{ij})$  and its vector of weights  $b = (b_i)$ . The method has *stage order*  $q$  if, with the nodes  $c_i = \sum_{j=1}^m a_{ij}$ ,

$$\sum_{j=1}^m a_{ij} c_j^{k-1} = \frac{c_i^k}{k} \quad (i = 1, \dots, m) \quad \text{for } k = 1, \dots, q.$$

We always assume that the quadrature formula with weights  $b_i$  and nodes  $c_i$  has at least the *quadrature order*  $q + 1$ :

$$\sum_{j=1}^m b_j c_j^{k-1} = \frac{1}{k} \quad \text{for } k = 1, \dots, q + 1.$$

In the following we consider Runge–Kutta methods that have the following important properties:

*Algebraic stability.* [1,4] The weights  $b_i$  are positive, and the matrix with entries  $b_i a_{ij} + b_j a_{ji} - b_i b_j$  is positive semidefinite.

*Coercivity.* [5] The Runge–Kutta coefficient matrix  $\mathcal{Q} = (a_{ij})$  is invertible, and there exist a positive diagonal matrix  $\mathcal{D} = \text{diag}(d_i)$  and  $\alpha > 0$  such that

$$v^T \mathcal{D} \mathcal{Q}^{-1} v \geq \alpha v^T \mathcal{D} v \quad \text{for all } v \in \mathbb{R}^m. \quad (4.4)$$

Important families of methods satisfying these properties are the Gauss and Radau IIA methods with an arbitrary number of stages  $m \geq 1$ ; see, e.g., [6] and [9, Chapter IV]. The  $m$ -stage Gauss and Radau IIA methods have stage order  $m$ , and have quadrature order  $2m$  and  $2m - 1$ , respectively.

## 4.2 Existence and uniqueness of the numerical solution

**Lemma 4.1** *Let conditions (K1)–(K3) hold and let the Runge–Kutta method satisfy the coercivity condition (4.4). For every  $R > 0$ , there exists  $\tau_R > 0$  such that the following holds: If  $u_n \in Y$  with  $\|u_n\| \leq R$ , then for stepsizes  $\tau \leq \tau_R$  the Runge–Kutta equations (4.1) and (4.3) have a unique solution in  $Y^m$  with*

$$\begin{aligned} |U_{ni}| &\leq C |u_n|, \\ \|U_{ni}\| &\leq C \|u_n\|, \end{aligned}$$

where  $C$  depends only on the Runge–Kutta coefficients.

*Proof* We may assume  $n = 0$  and write  $U_i, \dot{U}_i$  instead of  $U_{ni}, \dot{U}_{ni}$  for brevity. Similar to the proof of Lemma 3.3 the proof is based on constructing a contractive fixed-point iteration. Here we consider the map  $\Phi : (V_i)_{i=1}^m \mapsto (W_i)_{i=1}^m$  defined by the linear Runge–Kutta equations

$$W_i = u_0 + \tau \sum_{j=1}^m a_{ij} \dot{W}_j \quad (i = 1, \dots, m), \quad (4.5)$$

$$\dot{W}_i + A(V_i)W_i = 0. \quad (4.6)$$

We will show that, for sufficiently small stepsizes  $\tau$ , the map  $\Phi$  is well-defined and a contraction in the  $X^m$ -norm on

$$B_{cR} = \left\{ (V_i)_{i=1}^m \in Y^m : \sum_{i=1}^m d_i \|V_i\|^2 \leq (cR)^2 \right\},$$

where  $c$  is a constant depending only on the Runge–Kutta coefficients, which will be specified below. Again by [13, Lemma 7.3],  $B_{cR}$  is a closed set in  $X^m$ .

(i) We first prove that  $\Phi$  is a well-defined map from  $B_{cR}$  to itself for sufficiently small stepsizes  $\tau$ . For  $V = (V_i)_{i=1}^m$ , we write  $\mathcal{A}(V) = \text{diag}(A(V_i))$ . The equations (4.5)–(4.6) for  $W = (W_i)_{i=1}^m$  are then written compactly as

$$(I_m \otimes I + \tau(\mathcal{Q} \otimes I)\mathcal{A}(V))W = \mathbb{1} \otimes u_0$$

with  $\mathbb{1} = (1, \dots, 1)^T \in \mathbb{R}^m$ , or equivalently,

$$(\mathcal{Q}^{-1} \otimes I + \tau\mathcal{A}(V))W = (\mathcal{Q}^{-1}\mathbb{1}) \otimes u_0. \quad (4.7)$$

By conditions (K1) and (4.4), the linear operator

$$\mathcal{Q}^{-1} \otimes I - \alpha I_m \otimes I + \tau\mathcal{A}(V)$$

is  $m$ -accretive with respect to the inner product on  $X^m$  given by  $(W, \widetilde{W})_{\mathcal{D}} = \sum_{i=1}^m d_i (W_i, \widetilde{W}_i)$  with the corresponding norm  $|W|_{\mathcal{D}} = (W, W)_{\mathcal{D}}^{1/2}$ . Hence, equation (4.7) has a unique solution  $W \in D(\mathcal{A}(V))$ , and

$$|W|_{\mathcal{D}} \leq \alpha^{-1} |(\mathcal{Q}^{-1}\mathbb{1}) \otimes u_0|_{\mathcal{D}} \leq c_0 |u_0|,$$

where  $c_0$  depends only on the Runge–Kutta coefficients. We now recall condition (K2), which yields, with  $\mathcal{B}(V) = \text{diag}(B(V_i))$  and  $\mathcal{S} = I_m \otimes S$ ,

$$\mathcal{A}(V) = \mathcal{S}^{-1}\mathcal{A}(V)\mathcal{S} + \mathcal{S}^{-1}\mathcal{B}(V)\mathcal{S}.$$

Therefore,  $Z \in X^m$  is a solution of

$$((\mathcal{Q}^{-1} \otimes I) + \tau\mathcal{A}(V) + \tau\mathcal{B}(V))Z = (\mathcal{Q}^{-1}\mathbb{1}) \otimes (Su_0), \quad (4.8)$$

if and only if  $W = \mathcal{S}^{-1}Z \in Y^m$  solves (4.7). In view of (2.5), the  $X^m$  operator norm of  $\tau\mathcal{B}(V)$  is bounded by

$$|\tau\mathcal{B}(V)| \leq \tau M_{CR}^B,$$

where  $C = c/\min_i \sqrt{d_i}$ . For sufficiently small  $\tau$ , the operator norm of  $\tau\mathcal{B}(V)$  is therefore bounded by  $\alpha/2$ , and then equation (4.8) has a unique solution  $Z \in D(\mathcal{A}(V))$  and

$$\|W\|_{\mathcal{D}} = |Z|_{\mathcal{D}} \leq \frac{1}{\alpha - \tau C_1 M_{CR}^B} |(\mathcal{Q}^{-1}\mathbb{I}) \otimes Su_0|_{\mathcal{D}} \leq \frac{1}{\alpha - \tau M_{CR}^B} c_0 \|u_0\| \leq cR,$$

where  $\|W\|_{\mathcal{D}}^2 = \sum_{i=1}^m d_i \|W_i\|^2$  and  $c = 2c_0/\alpha$ .

(ii) Finally we show that  $\Phi : B_{cR} \rightarrow B_{cR}$  is a contraction with respect to the  $X^m$ -norm  $|\cdot|_{\mathcal{D}}$  for sufficiently small stepsizes  $\tau$ . Let  $W_i$  be defined by (4.5)–(4.6) and similarly  $\widetilde{W}_i$  by the same equations with  $V_i$  replaced by  $\widetilde{V}_i$ . We denote  $E_i = W_i - \widetilde{W}_i$  and  $\dot{E}_i = \dot{W}_i - \dot{\widetilde{W}}_i$  so that

$$\begin{aligned} E_i &= \tau \sum_{j=1}^m a_{ij} \dot{E}_j \\ \dot{E}_i + A(V_i)E_i &= -(A(V_i) - A(\widetilde{V}_i))\widetilde{W}_i. \end{aligned}$$

This is rewritten as

$$((\mathcal{Q}^{-1} \otimes I) + \tau\mathcal{A}(V))E = G := -\tau \left( (A(V_i) - A(\widetilde{V}_i))\widetilde{W}_i \right)_{i=1}^m.$$

We thus have, in view of the m-accretivity of  $\mathcal{A}(V)$ , of the Lipschitz bound (2.6) and the bound  $\|\widetilde{W}\|_{\mathcal{D}} \leq cR$ ,

$$|E|_{\mathcal{D}} \leq \alpha^{-1} |G|_{\mathcal{D}} \leq \alpha^{-1} \tau L_{CR}^A |V - \widetilde{V}|_{\mathcal{D}} cR.$$

This shows that  $\Phi$  is a contraction for sufficiently small  $\tau$ . The stated result then follows from the Banach fixed-point theorem.  $\square$

#### 4.3 Stability

**Lemma 4.2** *In addition to the conditions of Lemma 4.1, let the Runge–Kutta method be algebraically stable. For every  $R > 0$ , there exist  $\tau_R > 0$  and  $C_R > 0$  such that the following holds: If  $u_n \in Y$  with  $\|u_n\| \leq R$ , then for stepsizes  $\tau \leq \tau_R$  the Runge–Kutta equations (4.1)–(4.3) have a unique numerical solution  $u_{n+1} \in Y$  with*

$$\begin{aligned} |u_{n+1}| &\leq |u_n|, \\ \|u_{n+1}\| &\leq (1 + C_R \tau) \|u_n\|. \end{aligned}$$

*Proof* The proof follows closely the standard use of algebraic stability for contractive differential equations; see [1, 4] and, e.g., [9, Section IV.12]. We take again  $n = 0$  and write  $U_i$  for  $U_{ni}$ . By (4.2) we have

$$|u_1|^2 = |u_0|^2 + 2\tau \sum_{i=1}^m b_i(u_0, \dot{U}_i) + \tau^2 \sum_{i,j=1}^m b_i b_j (\dot{U}_i, \dot{U}_j).$$

Expressing  $u_0$  in the second term on the right-hand side by (4.1), we obtain

$$|u_1|^2 = |u_0|^2 + 2\tau \sum_{i=1}^m b_i(U_i, \dot{U}_i) - \tau^2 \sum_{i,j=1}^m (b_i a_{ij} + b_j a_{ji} - b_i b_j)(\dot{U}_i, \dot{U}_j).$$

By algebraic stability, we thus have

$$|u_1|^2 \leq |u_0|^2 + 2\tau \sum_{i=1}^m b_i(U_i, \dot{U}_i).$$

Since  $b_i > 0$  and  $(U_i, \dot{U}_i) = -(U_i, A(U_i)U_i) \leq 0$  by (K1), we obtain the bound  $|u_1|^2 \leq |u_0|^2$ .

For the bound in the  $Y$ -norm we obtain in the same way

$$|Su_1|^2 \leq |Su_0|^2 + 2\tau \sum_{i=1}^m b_i(SU_i, S\dot{U}_i).$$

Here we note, using subsequently (4.3), (K2), (K1) and Lemma 4.1,

$$\begin{aligned} (SU_i, S\dot{U}_i) &= -(SU_i, SA(U_i)S^{-1}SU_i) = -(SU_i, A(U_i)SU_i) - (SU_i, B(U_i)SU_i) \\ &\leq M_{CR}^B |SU_i|^2 = M_{CR}^B \|U_i\|^2 \leq M_{CR}^B (CR)^2 \|u_0\|^2, \end{aligned}$$

and the result follows.  $\square$

Suppose that  $u_n^* \in Y$  and  $U_{ni}^* \in Y$  (which will later be taken as the exact solution values  $u(t_n)$  and  $u(t_n + c_i\tau)$ ) solve (4.1)–(4.2) up to the defects  $d_{n+1} \in Y$  and  $D_{ni} \in Y$ , for  $0 \leq n\tau \leq T$ :

$$U_{ni}^* = u_n^* + \tau \sum_{j=1}^m a_{ij} \dot{U}_{nj}^* + D_{ni}, \quad i = 1, \dots, m, \quad (4.9)$$

$$u_{n+1}^* = u_n^* + \tau \sum_{i=1}^m b_i \dot{U}_{ni}^* + d_{n+1}, \quad (4.10)$$

where

$$\dot{U}_{ni}^* + A(U_{ni}^*)U_{ni}^* = 0, \quad i = 1, \dots, m.$$

**Lemma 4.3** *In the above situation, suppose that for  $0 \leq n\tau \leq T$  and for  $i = 1, \dots, m$ ,*

$$\|u_n^*\| \leq R, \quad \|U_{ni}^*\| \leq R \quad \text{and} \quad SU_{ni}^* \in Y \quad \text{with} \quad \|SU_{ni}^*\| \leq K.$$

*Then, there exist  $\bar{\tau} > 0$  and  $\delta > 0$ , which depend only on  $K$ ,  $R$ ,  $T$  and the coefficients of the Runge–Kutta method, such that for stepsizes  $\tau \leq \bar{\tau}$  and perturbations satisfying*

$$\|u_0 - u_0^*\|^2 + \tau \sum_{0 \leq n\tau \leq T} \left( \sum_{i=1}^m \|SD_{ni}\|^2 + \left\| \frac{d_{n+1}}{\tau} \right\|^2 + \|Sd_{n+1}\|^2 \right) \leq \delta^2,$$

the error satisfies, for  $0 \leq n\tau \leq T$ ,

$$\begin{aligned} \|u_n - u_n^*\|^2 &\leq C_Y \left( \|u_0 - u_0^*\|^2 + \tau \sum_{k=0}^{n-1} \left( \sum_{i=1}^m \|SD_{ki}\|^2 + \left\| \frac{d_{k+1}}{\tau} \right\|^2 + \|Sd_{k+1}\|^2 \right) \right), \\ |u_n - u_n^*|^2 &\leq C_X \left( |u_0 - u_0^*|^2 + \tau \sum_{k=0}^{n-1} \left( \sum_{i=1}^m \|D_{ki}\|^2 + \left| \frac{d_{k+1}}{\tau} \right|^2 + \|d_{k+1}\|^2 \right) \right), \end{aligned}$$

where  $C_Y$  depends only on  $K$ ,  $R$  and  $T$ , and  $C_X$  depends only on  $R$  and  $T$ .

*Proof* The proof is similar to those of Lemmas 2.1 and 3.2. We denote the errors by

$$e_n = u_n - u_n^*, \quad E_{ni} = U_{ni} - U_{ni}^*, \quad \dot{E}_{ni} = \dot{U}_{ni} - \dot{U}_{ni}^*.$$

We begin with  $n = 0$ , and we write  $E_i, U_i, U_i^*$  instead of  $E_{0i}, U_{0i}, U_{0i}^*$  and analogously for the corresponding quantities carrying a dot. By subtracting the original and perturbed Runge–Kutta equations we obtain

$$E_i = e_0 + \tau \sum_{j=1}^m a_{ij} \dot{E}_j - D_i, \quad i = 1, \dots, m, \quad (4.11)$$

$$e_1 = e_0 + \tau \sum_{i=1}^m b_i \dot{E}_i - d_1, \quad (4.12)$$

where

$$\dot{E}_i + A(U_i)E_i = -(A(U_i) - A(U_i^*))U_i^*, \quad i = 1, \dots, m. \quad (4.13)$$

By condition (K2), the latter equation is equivalent to

$$\begin{aligned} S\dot{E}_i + A(U_i)SE_i + B(U_i)SE_i &= -(A(U_i) - A(U_i^*))SU_i^* \\ &\quad - (B(U_i) - B(U_i^*))SU_i^*. \end{aligned} \quad (4.14)$$

As in the proof of Lemma 4.2, using (4.11)–(4.12) and algebraic stability we obtain

$$\|e_1\|^2 - \|e_0\|^2 \leq 2\tau \sum_{i=1}^m (SE_i + SD_i, S\dot{E}_i) - 2(S e_0 + \tau \sum_{i=1}^m b_i S\dot{E}_i, Sd_1) + \|d_1\|^2.$$

We estimate the first two terms on the right-hand side separately. We write

$$\begin{aligned} (SE_i + SD_i, S\dot{E}_i) &= -(SE_i, A(U_i)SE_i) - (A(U_i)^*SD_i, SE_i) - (SE_i + SD_i, B(U_i)SE_i) \\ &\quad - (SE_i + SD_i, (A(U_i) - A(U_i^*))SU_i^*) - (SE_i + SD_i, (B(U_i) - B(U_i^*))SU_i^*), \end{aligned}$$

and note that the adjoint operator  $A(y)^*$  is bounded like  $A(y)$  for  $\|y\| \leq CR$ : using that  $Y$  is dense in  $X$ , and condition (K2) and recalling that  $S$  is an isometry between  $Y$  and  $X$ ,

$$\begin{aligned} |A(y)^*w| &= \sup_{0 \neq v \in Y} \frac{(A(y)^*w, v)}{|v|} = \sup_{0 \neq v \in Y} \frac{(w, A(y)v)}{|v|} \\ &= \sup_{0 \neq v \in Y} \frac{(Sw, S^{-1}A(y)SS^{-1}v)}{|v|} = \sup_{0 \neq v \in Y} \frac{(Sw, A(y)S^{-1}v + B(y)S^{-1}v)}{|v|} \\ &\leq \sup_{0 \neq v \in Y} \frac{|Sw| \cdot (M_{CR}^A + M_{CR}^B)\|S^{-1}v\|}{|v|} = \|w\| \cdot (M_{CR}^A + M_{CR}^B). \end{aligned}$$

Using the relation (4.14) and the accretivity (2.2), the bounds (2.5)–(2.9), we therefore obtain, as long as  $\|U_i\| \leq CR$  ( $i = 1, \dots, m$ ),

$$\begin{aligned} (SE_i + SD_i, S\dot{E}_i) &\leq (M_{CR}^A + M_{CR}^B)\|SD_i\|\|E_i\| \\ &\quad + (M_{CR}^B + L_{CR}^AK + L_{CR}^BR)\|E_i + D_i\|\|E_i\| \\ &\leq C_{K,R}\|E_i\|^2 + C_{K,R}\|SD_i\|^2, \end{aligned}$$

where we also used the norm relation  $\|D_i\| = |SD_i| \leq \|SD_i\|$ .

The other term is estimated similarly:

$$(Se_0 + \tau \sum_{i=1}^m b_i S\dot{E}_i, Sd_1) \leq \|e_0\|\|d_1\| + \tau \sum_{i=1}^m b_i (S\dot{E}_i, Sd_1),$$

where the terms in the sum are bounded by

$$\begin{aligned} (S\dot{E}_i, Sd_1) &\leq (M_{CR}^A + M_{CR}^B)\|E_i\|\|Sd_1\| + (L_{CR}^AK + L_{CR}^BR)\|E_i\|\|d_1\| \\ &\leq \|E_i\|^2 + C_{K,R}\|Sd_1\|^2. \end{aligned}$$

By combining these estimates we obtain

$$\begin{aligned} (Se_0 + \tau \sum_{i=1}^m b_i S\dot{E}_i, Sd_1) &\leq \tau\|e_0\|^2 + \tau c_0 \sum_{i=1}^m \|E_i\|^2 \\ &\quad + \tau C_{K,R}\|Sd_1\|^2 + \tau C_{K,R} \left\| \frac{d_1}{\tau} \right\|^2. \end{aligned}$$

Altogether, we have

$$\begin{aligned} \|e_1\|^2 - \|e_0\|^2 &\leq \tau\|e_0\|^2 + \tau C_{K,R} \sum_{i=1}^m \|E_i\|^2 \\ &\quad + \tau C_{K,R} \sum_{i=1}^m \|SD_i\|^2 + \tau C_{K,R} \left\| \frac{d_1}{\tau} \right\|^2 + \tau C_{K,R}\|Sd_1\|^2. \end{aligned}$$

To estimate the terms  $\|E_i\|^2$ , we use the coercivity property (4.4) of the Runge–Kutta method. We use the notations of the proof of Lemma 4.1 and  $E =$



$(E_1, \dots, E_m)^T$ ,  $\dot{E} = (\dot{E}_1, \dots, \dot{E}_m)^T$  and  $D = (D_1, \dots, D_m)^T$ . We thus rewrite (4.1) as

$$E = \mathbb{I} \otimes e_0 + \tau(\mathcal{Q} \otimes I)\dot{E} - D.$$

We multiply both sides by  $\mathcal{D}\mathcal{Q}^{-1} \otimes S$ , use (4.14) and (K2), and then take the inner product with  $SE$ , where again  $\mathcal{S} = I_m \otimes S$ . Using similar estimates as above we obtain

$$\begin{aligned} (SE, (\mathcal{D}\mathcal{Q}^{-1} \otimes I)SE) &= \tau(SE, \mathcal{S}\dot{E})_{\mathcal{D}} + (SE, (\mathcal{D}\mathcal{Q}^{-1} \otimes I)(\mathbb{I} \otimes Se_0 - SD)) \\ &\leq \tau(M_{CR}^B + L_{CR}^A mK + L_{CR}^B mR) \|E\|_{\mathcal{D}}^2 \\ &\quad + c_0 \|E\|_{\mathcal{D}} (\|e_0\| + \|D\|) \\ &\leq \tau C_{K,R} \|E\|_{\mathcal{D}}^2 + c_0 \|E\|_{\mathcal{D}} (\|e_0\| + \|D\|), \end{aligned}$$

where the constant  $c_0$  only depends on the method. Using the coercivity of the Runge–Kutta method on the left-hand side, an absorption (by choosing the stepsize to satisfy  $\tau C_{K,R} \leq \alpha/2$ ) and Young's inequality for the right-hand side yields the bound

$$\sum_{i=1}^m \|E_i\|^2 \leq c \left( \|e_0\|^2 + \sum_{i=1}^m \|D_i\|^2 \right).$$

Finally, combining all estimates we obtain

$$\begin{aligned} \|e_1\|^2 - \|e_0\|^2 &\leq \tau(1 + C_{K,R}c) \|e_0\|^2 \\ &\quad + \tau C_{K,R} \sum_{i=1}^m \|SD_i\|^2 + \tau C_{K,R} \left\| \frac{d_1}{\tau} \right\|^2 + \tau C_R \|Sd_1\|^2. \end{aligned}$$

The analogous estimate for  $\|e_{n+1}\|^2 - \|e_n\|^2$  holds for all  $n$  as long as  $\|U_{ni}\| \leq CR$ . Summing over  $n$  and applying the discrete Gronwall inequality, we obtain the stated error bound in the  $Y$ -norm. Choosing  $\delta$  so small that  $C_Y \delta \leq R$ , the condition  $\|U_{ni}\| \leq CR$  then remains satisfied for  $0 \leq n\tau \leq T$ . The  $X$ -norm error bound is obtained analogously, using (2.2) and (2.6).  $\square$

#### 4.4 Convergence with the stage order plus 1

Using  $u_n^* = u(t_n)$  and  $U_{ni}^* = u(t_n + c_i\tau)$  in Lemma 4.3, we obtain the following error bound.

**Theorem 4.1** *Let the conditions (K1)–(K4) be satisfied, and suppose that the solution  $u$  of (2.1) has the regularity  $u \in C^{q+2}([0, T], Y)$  with  $Su \in C^{q+1}([0, T], Y)$ . Then, there exists  $\bar{\tau} > 0$  such that for stepsizes  $0 < \tau \leq \bar{\tau}$ , the errors of an algebraically stable and coercive Runge–Kutta method with stage order  $q$  and quadrature order at least  $q + 1$  are bounded by*

$$\|u_n - u(t_n)\| \leq C\tau^{q+1},$$

where  $C$  is independent of  $n$  and  $\tau$  with  $0 \leq n\tau \leq T$ .

*Proof* With the choice  $u_n^* = u(t_n)$  and  $U_{ni}^* = u(t_n + c_i\tau)$  in Lemma 4.3, the defects  $D_{ni}$  and  $d_{n+1}$  in (4.9) and (4.10) are just quadrature errors:

$$D_{ni} = \tau^q \int_{t_n}^{t_{n+1}} \kappa_i\left(\frac{t-t_n}{\tau}\right) u^{(q+1)}(t) dt,$$

$$d_{n+1} = \tau^{q+1} \int_{t_n}^{t_{n+1}} \kappa\left(\frac{t-t_n}{\tau}\right) u^{(q+2)}(t) dt = -\tau^q \int_{t_n}^{t_{n+1}} \kappa'\left(\frac{t-t_n}{\tau}\right) u^{(q+1)}(t) dt$$

with real-valued, bounded Peano kernels  $\kappa_i$  and  $\kappa$ . The result then follows from Lemma 4.3.  $\square$

#### 4.5 Convergence with the classical order

A Runge–Kutta method has *classical order*  $p$  if the local error (i.e., the error after one step starting from the exact solution) is of size  $\mathcal{O}(\tau^{p+1})$  whenever the method is applied to an ordinary differential equation  $\dot{y} = f(y)$  in  $\mathbb{R}^n$  with an arbitrarily differentiable function  $f$ . We recall that the classical order of the  $m$ -stage Gauss and Radau IIA methods is  $2m$  and  $2m - 1$ , respectively, whereas the stage order of these methods is  $m$ ; see [9, Chapter IV].

We now show that for the quasi-linear problem (2.1) we can retain the classical order under additional regularity conditions. The first such condition is a generalization of condition (K2):

For  $k = 1, \dots, p - q$  and for every  $y \in Y$  with  $\|y\| \leq R$ ,

$$S^k A(y) S^{-k} = A(y) + B_k(y), \quad (4.15)$$

where  $B_k(y)$  is  $Y$ -locally uniformly bounded on  $X$ :

$$|B_k(y)v| \leq M_{k,R} |v| \quad \text{for all } v \in X. \quad (4.16)$$

With  $L(Y, X)$  denoting the Banach space of bounded linear operators from  $Y$  to  $X$  (and analogously  $L(X, X)$ ), we further suppose that the operators  $A$  and  $B$  of (2.4) satisfy the following:

$A(\cdot) : y \in Y \mapsto A(y) \in L(Y, X)$  and  $B_k(\cdot) : y \in Y \mapsto B_k(y) \in L(X, X)$  are arbitrarily differentiable, and for any  $R > 0$ , their derivatives up to any fixed order are uniformly bounded for  $\|y\| \leq R$ .

In the presence of a semilinear term  $f(y)$  in (2.1) (which we have discarded in this section), a similar  $Y$ -locally uniform differentiability condition is required for  $f$ . We note that the above conditions are satisfied in all the examples of [13].

The following theorem can be viewed as an extension of the full-order error bounds for *linear* evolution equations in [3, 15, 16] to the quasi-linear case studied here.

**Theorem 4.2** *Let (K1)–(K4) and the above conditions be satisfied, and suppose that the solution  $u$  of (2.1) has the regularity  $u \in C^{p+1}([0, T], Y)$  with  $S^k u \in C^{p+1-k}([0, T], Y)$  for  $k = 1, \dots, p - q$ . Then, there exists  $\bar{\tau} > 0$  such that for stepsizes  $0 < \tau \leq \bar{\tau}$ , the errors of an algebraically stable and coercive Runge–Kutta method with stage order  $q$  and classical order  $p$  (with  $2q \geq p$ ) are bounded by*

$$\|u_n - u(t_n)\| \leq C\tau^p,$$

where  $C$  is independent of  $n$  and  $\tau$  with  $0 \leq n\tau \leq T$ .

*Remark* The condition  $2q \geq p$  simplifies the proof and is satisfied for the Gauss and Radau IIA methods, which are arguably the most interesting classes of implicit Runge–Kutta methods. We expect, however, that this condition can be dropped.

*Proof* (a) Let us first show how we get from order of convergence  $q + 1$  to order  $q + 2$ . We start by taking as  $U_{ni}^*$  in (4.9) the exact solution value at  $t_n + c_i\tau$ . In the following we can again take  $n = 0$  and drop the dependence on  $n$  in the notation. We then modify the reference internal stages by setting

$$U_i^{[q+2]} = U_i^* - D_i$$

and  $\dot{U}_i^{[q+2]} = -A(U_i^{[q+2]})U_i^{[q+2]}$ . The modified defect  $D_i^{[q+2]}$  in the Runge–Kutta equations,

$$U_i^{[q+2]} = u(t_0) + \tau \sum_{j=1}^m a_{ij} \dot{U}_j^{[q+2]} + D_i^{[q+2]},$$

is then

$$D_i^{[q+2]} = \tau \sum_{j=1}^m a_{ij} (A(U_j^*)U_j^* - A(U_j^* - D_j)(U_j^* - D_j)),$$

which by (K1)–(K3), by the Peano kernel formula for  $D_i$  in the proof of Theorem 4.1 and the regularity assumption for the exact solution is bounded by

$$\begin{aligned} \|D_i^{[q+2]}\| &= |SD_i^{[q+2]}| \\ &\leq \tau \sum_{j=1}^m |a_{ij}| |S(A(U_j^*)D_j + (A(U_j^*) - A(U_j^* - D_j))(U_j^* - D_j))| \\ &\leq \tau \sum_{j=1}^m |a_{ij}| |A(U_j^*)SD_j + B(U_j^*)SD_j + (A(U_j^*) - A(U_j^* - D_j))(SU_j^* - SD_j) \\ &\quad + (B(U_j^*) - B(U_j^* - D_j))(SU_j^* - SD_j)| \\ &\leq c\tau \max_{1 \leq j \leq m} \|SD_j\| \leq C\tau^{q+2}. \end{aligned}$$

The defect in

$$u(t_1) = u(t_0) + \tau \sum_{i=1}^m b_i \dot{U}_i^{[q+2]} + d_1^{[q+2]}$$

is then

$$d_1^{[q+2]} = d_1 + \tau \sum_{i=1}^m b_i (A(U_i^*)U_i^* - A(U_i^* - D_i)(U_i^* - D_i)),$$

where we know already that

$$\|d_1\| \leq C\tau^{p+1} \leq C\tau^{q+3}$$

in the case of interest where  $p \geq q + 2$ . The more challenging term is

$$\begin{aligned} & \tau \sum_{i=1}^m b_i (A(U_i^*)U_i^* - A(U_i^* - D_i)(U_i^* - D_i)) \\ &= \tau \sum_{i=1}^m b_i \left( A(U_i^*)D_i - \int_0^1 A'(U_i^* - \theta D_i)[D_i](U_i^* - D_i) d\theta \right). \end{aligned}$$

This differs by  $\mathcal{O}(\tau^{q+3})$  in the  $Y$ -norm from

$$\tau \sum_{i=1}^m b_i (A(u(t_0))D_i - A'(u(t_0))[D_i]u(t_0)) = 0,$$

because we have the quadrature error

$$D_i = \tau^{q+1} \left( \sum_{j=1}^m a_{ij} c_j^q - \frac{c_i^{q+1}}{q+1} \right) u^{(q+1)}(t_0) + \mathcal{O}(\tau^{q+2}),$$

and the order conditions for the  $p$ th-order Runge–Kutta method (see [2, 8]) yield

$$\sum_{i=1}^m b_i \left( \sum_{j=1}^m a_{ij} c_j^q - \frac{c_i^{q+1}}{q+1} \right) = 0.$$

Hence,

$$\|d_1^{[q+2]}\| \leq C\tau^{q+3}.$$

Lemma 4.3 used with  $U_i^{[q+2]}$  in the role of  $U_i^*$  then yields the result for  $p = q+2$ .

(b) The above procedure for modifying the reference internal stages can be repeated. In the next step we set

$$U_i^{[q+3]} = U_i^{[q+2]} - D_i^{[q+2]}$$

and so on we iterate up to  $U_i^{[p]}$ . Under the given regularity conditions we then obtain defects

$$U_i^{[p]} = u(t_0) + \tau \sum_{j=1}^m a_{ij} \dot{U}_j^{[p]} + D_i^{[p]}$$

with

$$\|D_i^{[p]}\| \leq C\tau^p,$$

gaining a factor  $\tau$  at the expense of an application of  $S$  to the previous defect in every iteration. The defect  $d_1^{[p]}$  in

$$u(t_1) = u(t_0) + \tau \sum_{i=1}^m b_i \dot{U}_i^{[p]} + d_1^{[p]}$$

then becomes a more complicated expression than before, but the key observation is that it can be Taylor-expanded into terms of the form

$$\tau^k \sum_{i, j_1, \dots, j_r=1}^m b_i c_i^{\ell_0} a_{ij_1} c_{j_1}^{\ell_1} \dots a_{j_{r-1}j_r} c_{j_r}^{\ell_r} \left( \sum_{j=1}^m a_{j_r j} c_j^{s-1} - \frac{c_{j_r}^s}{s} \right)$$

multiplied with an expression depending on the solution  $u$  and its derivatives evaluated at  $t_0$ . We omit the details. For  $k \leq p$  these terms all vanish by the order conditions of the Runge–Kutta method [2, 8]. In this way we obtain

$$\|d_1^{[p]}\| \leq C\tau^{p+1} \quad \text{and} \quad \|Sd_1^{[p]}\| \leq C\tau^p,$$

and the result then follows again by Lemma 4.3 with  $U_i^{[p]}$  in the role of  $U_i^*$ .  $\square$

## Acknowledgement

This work was supported by Deutsche Forschungsgemeinschaft, SFB 1173.

## References

1. K. Burrage and J. C. Butcher, *Stability criteria for implicit Runge–Kutta methods*, SIAM Journal on Numerical Analysis **16** (1979), 46–57.
2. J. C. Butcher, *Numerical methods for ordinary differential equations*, John Wiley & Sons, 2008.
3. M. Crouzeix, *Sur l'approximation des équations différentielles opérationnelles linéaires par des méthodes de Runge–Kutta*, Thèse d'Etat, Univ. Paris 6, 1975.
4. ———, *Sur la B-stabilité des méthodes de Runge–Kutta*, Numerische Mathematik **32** (1979), 75–82.
5. M. Crouzeix and P.-A. Raviart, *Approximation des problèmes d'évolution*, Lecture Notes, Univ. Rennes, 1980.
6. K. Dekker and J. Verwer, *Stability of Runge–Kutta methods for stiff nonlinear differential equations*, Elsevier, 1984.
7. W. Dörfler, H. Gerner, and R. Schnaubelt, *Local well-posedness of a quasilinear wave equation*, Applicable Analysis (2015), 1–14.
8. E. Hairer, S. P. Nørsett, and G. Wanner, *Solving ordinary differential equations I. Nonstiff differential equations*, Second ed., Springer, Berlin, 1993.
9. E. Hairer and G. Wanner, *Solving ordinary differential equations II. Stiff and differential-algebraic problems*, Second ed., Springer, Berlin, 1996.
10. M. Hochbruck and T. Pažur, *Error analysis of implicit Euler methods for quasilinear hyperbolic evolution equations*, Numerische Mathematik (2016), 1–23.
11. T. J. R. Hughes, T. Kato, and J. E. Marsden, *Well-posed quasi-linear second-order hyperbolic systems with applications to nonlinear elastodynamics and general relativity*, Archive for Rational Mechanics and Analysis **63** (1977), 273–294.

12. T. Kato, *The Cauchy problem for quasi-linear symmetric hyperbolic systems*, Archive for Rational Mechanics and Analysis **58** (1975), 181–205.
13. ———, *Quasi-linear equations of evolution, with applications to partial differential equations*, Spectral theory and differential equations (Proc. Sympos., Dundee, 1974; dedicated to Konrad Jörgens), Lecture Notes in Math., Vol. 448, Springer, Berlin, 1975, pp. 25–70.
14. ———, *Perturbation theory for linear operators*, Classics in Mathematics, Springer-Verlag, Berlin, 1995, Reprint of the 1980 edition.
15. C. Lubich and A. Ostermann, *Interior estimates for time discretizations of parabolic equations*, Applied numerical mathematics **18** (1995), 241–251.
16. D. Mansour, *Gauss–Runge–Kutta time discretization of wave equations on evolving surfaces*, Numerische Mathematik **129** (2015), 21–53.
17. D. Müller, *Well-posedness for a general class of quasilinear evolution equations — with applications to Maxwell’s equations*, Ph.D. thesis, KIT, <http://www.math.kit.edu/iana3/dmueller/media/thesis.pdf>, 2014.
18. A. Pazy, *Semigroups of linear operators and applications to partial differential equations*, Applied Mathematical Sciences, vol. 44, Springer-Verlag, New York, 1983.